# 融合深层差异特征的 RGB-T 巢式语义分割网络 *

袁浩宾，赵　涛，钟羽中

(四川大学 电气工程学院, 成都 610065)

**摘　要：** 针对现存可见光-红外(RGB-T)图像语义分割模型分割性能不高的问题，提出一种基于深层差异特征互补融合的巢式分割网络。具体来说，网络的编码部分和解码部分通过多级稠密中间路径相连形成一个嵌套形式的结构，编码器的深浅特征通过多级路径供解码器实现密集的多尺度特征复用，另一方面多模态深层特征通过特征差异性融合策略增强其语义表达能力。实验结果表明，所提网络在 MFNet 数据集上实现了 65.8%的平均准确率和 54.7%的平均交并比，与其他先进 RGB-T 分割模型相比，具有更优越的分割能力。

**关键词：** RGB-T 语义分割；巢式网络；特征复用；融合策略

**中图分类号：** TP391.41　　doi: 10.19734/j.issn.1001-3695.2022.03.0083

## Nested semantic segmentation network fusing deep difference features

Yuan Haobin, Zhao Tao, Zhong Yuzhong

(*College of Electrical Engineering*, *Sichuan University*, *Chengdu* 610065,*China)*

**Abstract:** Considering the existing visible-infrared image (RGB-T) semantic segmentation models have limitations in segmentation performance, a nested semantic segmentation network fusing deep difference features is proposed. Specifically, the encoding part and the decoding part of the network are connected by a multi-level dense intermediate path to form a nested structure, and encoder features at various levels achieve densely repeated utilization via multi-stage path while the multi-modal deep feature enhances its semantic expressiveness by the feature differential fusion strategy. The comparison experiments show that the proposed network achieves an average accuracy of 65.8% and an average intersection over union of 54.7% on the MFNet dataset. Compared with other state-of-the-art RGB-T segmentation models, it has better segmentation ability.

**Key words:** RGB-T semantic segmentation; nested network; feature reultilization; fusion strategy



## 0　引言

语义分割旨在从像素级层面上为图像划分所属类别，在自动驾驶[1]、医疗分析[2]和机器人定位[3]等领域具有广泛的应用空间。受可见光传感器成像机制所限[4]，当前主流的 RGB 分割模型在浓雾和暗光等条件下存在不可避免地性能退化[5]。得益于红外传感器捕获热辐射信息，红外图像可以有效补偿劣势环境下 RGB 图像中的缺漏信息[6]，因而融合这两种模态图像进行场景表征具有更强的健壮性。

RGB-T 语义分割近几年备受研究者青睐。MFNet[7]是首个用于自动驾驶的 RGB-T 实时语义分割网络，该模型受 FuseNet 架构[8]启发，由两个对称的低参数编码器和单个解码器组成，编码器末两层通过微型下采样感知模块捕获更大感受野的多尺度特征。RTFNet[9]利用 ResNet[10]作为两个编码器的骨干结构整合 RGB 和红外图像信息，解码部分通过两种类型的上采样模块逐层渐进式的恢复分辨率和重构特征。Xu 等人[11]将编码器改进为带空洞卷积运算的 ResNet 网络以提高对小目标的检测，并设计了一个共注意力机制模块来融合提取的多模态特征。Guo 等人[12]关注多尺度信息的利用，提出了一个辅助解码模块来接收编码器的各级特征，这种跨尺度特征传递的方式实现了更灵活的上下文信息融合。

这些研究对 RGB-T 语义分割作出了不同层面的贡献，但有以下挑战存在改进空间。首先，仅仅依赖深层特征单向传递到顺序相连的解码层会因编码下采样过程而丢失图像的部分边缘细节信息[9,11]，而通过跳跃连接在解码端复用同尺度编码特征一定程度上缓解了该问题[7]，但深浅特征利用方式仍不够充分。此外，编码器在特征融合阶段未充分考虑到 RGB 和红外图像的特征模态差异存在，例如在黑夜环境下，红外图像包含 RGB 图像不能感知到的信息内容，通过简单相加[9]和在通道层面拼接[7]，某些情况下会对易辨识的特征造成对冲作用，削弱优势

特征的编码响应，尤其对高维特征影响更为突出，而采取基于 Softmax 算子的共注意力[11]进行融合的方式缺乏学习能力。

为更加充分复用各级编码特征和减少模特差异对高维特征的融合影响，本文提出了一种融合 RGB 和红外图像深层差异特征的 RGB-T 巢式语义分割网络。其贡献在于：

a)编码器深浅特征密集复用方式。编码器和解码器通过多级中间路径相接，来自不同层次的尺度相异的编码特征通过叠加的方式整合并馈送到解码端，解码层能利用到更多的多尺度特征信息帮助语义划分。

b)深层特征融合策略。在深层特征融合阶段，针对 RGB 和红外图像性质的差异性，设计一种特征差异性融合策略完成两种模特图像的互补特征提取，从而实现多模态特征更好的信息融合，深层高维抽象特征的语义表征能力因而得到增强。

## 1　巢式语义分割网络

巢式连接架构最早由 Zhou 等人[13]在医学图像分割任务中提出，基于不同层次特征对尺寸大小不同的目标对象表现出不同敏感度这一事实，他们将 U-Net 网络[14]中的长跳跃连

---

**收稿日期：** 2022-03-14；**修回日期：** 2022-04-20　**基金项目：** 国家重点研发计划资助项目(2018YFB1307401)

**作者简介：** 袁浩宾(1997-)，男，四川达州人，硕士研究生，主要研究方向为图像语义分割和视觉 SLAM(yyyhya28@163.com)；赵涛，男，副教授，博导，博士，主要研究方向为智能机器人控制和模糊控制；钟羽中，女，助理研究员，硕导，博士，主要研究方向为计算机视觉.

接替换为上采样和长短跳跃组合的嵌套巢式连接。图 1 为巢式连接的框架结构。

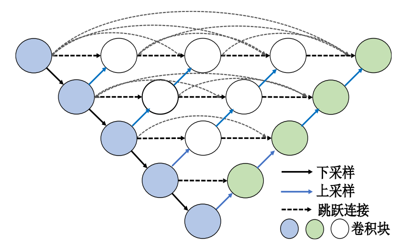



图 1　巢式连接结构

Fig. 1　The architecture of nest connection

在巢式结构中，深浅层的编码特征通过上采样和稠密连接在通道上进行密集拼接和复用，各个不同层次的特征因此得到有效整合。受此启发，本文将巢式结构引入到 RGB-T 语义分割任务中，构建能够充分整合所有尺度特征信息的 RGB-T 分割网络。如图 2 所示，所提分割模型包含两个结构一致的编码器和一个解码器，左侧双编码器逐层降采样提取深浅特征，右侧解码器渐进式的重构特征，编码部分和解码部分通过稠密连接的多级中间过渡层相连，整体上形成一个嵌套形式的巢式网络。相比于现存 RGB-T 分割网络，密集的中间信息流通渠道使各级语义特征信息得到有效保留。

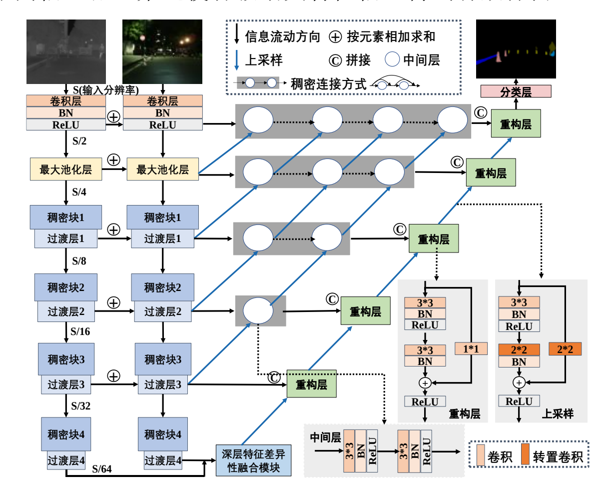



图 2　RGB-T 巢式语义分割网络

Fig. 2　RGB-T nested semantic segmentation network

## 1.1　深浅特征编码多级复用

众多 RGB-T 分割模型采用 ResNet 作为骨干结构，考虑到 DenseNet[15]拥有更为密集的信息传播途径且参数量更少，本文编码器的骨干网络使用 DenseNet 框架。同时为保留更多的原始空间信息和加强编码器内部结构的统一性，DenseNet 的分类层被舍弃，并在第 4 个稠密块之后增补了与其他过渡层结构一致的过渡层。因而，编码器可以划分为初始卷积层、最大池化层和 4 个由稠密块和过渡层组成的稠密特征单元，其中稠密块保持特征图的分辨率不变，剩余部分实现 2 倍率的下采样。考虑到红外图像为单通道灰度图，红外编码器的初始卷积层的输入通道数修改为 1。对于前 5 个下采样过程，RGB 和红外信息通过按元素相加的方式进行特征融合，对于末尾下采样阶段提取的深层高维特征，通过特征差异性融合策略完成融合。

在所提模型中，各层融合特征通过上采样和中间层进行信息多级回流，回流特征和前一融合特征的输出密集的堆叠在一起，并传递至对应层级的重构层输入端。和仅使用长跳跃连接相比，网络编码层和解码层间的语义鸿沟能够通过中间层得到缓解。如图 1 所示，上采样单元类似残差结构，通过转置卷积实现特征分辨率倍增，中间层由两个级联的卷积层构成，避免了单个卷积的非线性特征提取能力的缺乏。

## 1.2　深层差异特征互补融合

末尾稠密特征单元传递深层信息的渠道仅有一条，在进行解码重构时存在这样一个挑战：深层网络捕获到小尺度等较困难目标的梯度信息较小，此时 RGB 和红外特征表现出更高维度的抽象语义性，特别是在不利光照环境成像下，RGB 图像携带的盲区信息会使其深层特征更难以学习，此时结合红外信息应当更多的专注在能够弥补双方的弱势特征区域。鉴于 RGB 和红外图像成像原理具有差异性，通过在像素层面上构建双模态图像特征的差异性，提出了一种基于特征差异性的互补融合策略，用以增强深层特征的语义表达。

如图 3 所示，差异性融合模块的输入为 RGB 和红外特征图，在 RGB 深层特征编码阶段，双模态特征首先经由卷积运算得到通道压缩后的特征映射矩阵 $Q_r$ 和 $K_t$，两个矩阵在空间尺度展开后进行如下运算获取模态特征差异性权重矩阵：

$$W_{rt} = 1 - softmax(Q_r K_t^T) \tag{1}$$

特征图在像素级层面表现为数值向量矩阵，$Q_r$ 和 $K_t^T$ 相乘反映了 RGB 和红外特征的特征相关度。softmax 归一化运算保证相关度矩阵为反映公共特征在全局位置上的权重系数，因而模态特征的差异性可通过其和 1 的补数表示。接着，RGB 特征图的线性变换矩阵 $W_{rt}$ 和 $V_r$ 进行加权处理获取 RGB 特征图的互补特征：

$$Feature_{rt} = W_{rt} V_r \tag{2}$$

同样在红外特征编码阶段通过上述处理获取红外特征图的互补特征 $Feature_{tr}$。最后两个互补特征同输入双模态特征相加实现深层特征互补融合增强。

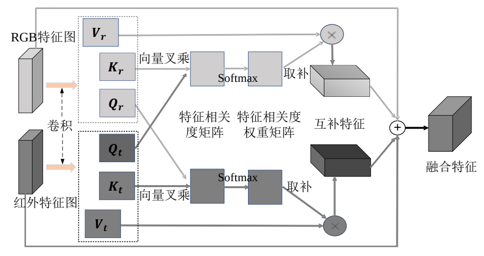



图 3　深层差异特征融合策略

Fig. 3　Deep differential feature fusion strategy

## 1.3　特征解码器

解码器依靠接收的编码特征进行重构，获取稠密的像素预测。所提网络的解码器包括上采样、重构层和分类层，详见图 1，其中分类层由单个卷积层和双线性插值运算构成，和上采样模块功能一致，实现倍增的特征图分辨率恢复，并完成像素信息的语义归类。分类层的卷积输出通道数量设置为语义类别总数。为增强网络梯度传播，重构层通过两个顺序相连的卷积层和一个残差路径上的 1*1 卷积构成残差结构。由于每个重构层接收来自同尺度和低尺度的堆叠特征，重构层的第一个卷积和残差层保证特征图通道数缩减为相同层级的编码层的输出通道数，第二个卷积维持特征图分辨率和通道数不变。网络所有的卷积层后紧跟一个批归一化和 RuLU 层。总的来说，解码器可划分为 5 个由上采样模块和重构层组成的重构单元和 1 个分类层，多级别深浅特征复用路径有效帮助语义预测，渐进形式的特征尺度恢复保证了解码器和编码器在结构上的对称性。

鉴于 DenseNet 拥有卷积层数相异的变体:DenseNet-121、DenseNet-169、DenseNet-201 和 DenseNet-161，前三个架构的特征通道增长率为 32，末尾的为 48，它们的参数复杂度依次递增。在采用不同变体结构时，各个降采样阶段的特征输出通道与相应变体对齐，解码器的重构单元的输入特征通道数也相应变动。

## 1.4　损失函数

损失函数同网络拟合方向和收敛速度密切相关。通常语

义分割领域采用交叉熵完成训练：

$$L_{CE} = -\sum_{c=1}^{M} y_c \log(p_c) \tag{3}$$

其中，M 为类别数，$y_c$ 和 $p_c$ 分别表示目标图像类别划分属于 c 的真值标签向量和预测概率图。考虑到图像的各尺度目标分布不可能完全均衡，交叉熵损失不能很好的平衡这种样本差异，本文额外引入改进的 DiceLoss[16]项增强网络学习能力：

$$L_{dl} = 1 - \frac{2\sum_{i}^{N} p_i g_i}{\sum_{i}^{N} p_i^2 + \sum_{i}^{N} g_i^2} \tag{4}$$

其中，$p_i$ 和 $g_i$ 分别表示目标图像的像素域 N 内的第 i 个像素的二进制预测值和二进制真实标签值。因而网络的总损失表示为

$$L_{total} = \frac{1}{2}(L_{CE} + L_{dl}) \tag{5}$$

由于两个损失项的值域具有相同数量级，它们各自占有一半的权重。这两项共同引导网络学习，弥补了使用单一交叉熵损失项的不足。

## 2　实验与分析

### 2.1　数据集与训练细节

MFNet 发布了首个基于像素级语义标注的 RGB-T 城市道路场景图像数据集，其中白天和夜晚采集的 RGB-红外图像对各有 820 对和 749 对，图像分辨率统一为 480×640 大小。该数据集手工标记了行车道路上的 9 个语义类：汽车(Car)、行人(Person)、单车(Bike)、车道线(Curve)、停车位(Car Stop)、护栏(Guardrail)、色锥(Color Cone)、路面凸起物(Bump)和未标记背景区(Unlabelled)，每个类别的像素数量极其不均衡，尤以停车位和护栏类为甚。本文遵循原始数据集的划分方案，训练集和验证集的图像数量占比为 2：1,其中昼夜图片对半，剩余 393 对图像用作测试集。

网络模型部署在 PyTorch 框架上，使用随机梯度下降(SGD)策略作为优化器。网络各层通过 Xavier 方案[17]进行权重初始化，学习率从 $1\times10^{-2}$ 开始按 0.95 的衰减权重逐个 epoch 进行指数衰减。输入图像通过像素归一化至[0,1]区间，并且在每个 epoch 前随机翻转处理以预防网络过拟合。BatchSize 根据骨干网络变体结构相应调整，DenseNet-161 设为 2，DenseNet-201 和 DenseNet-169 设为 4，DenseNet-121 设为 6。所有训练和测试过程均在一台配备 24G 显存的 NVIDIA GeForce RTX 3090 GPU、32GB 内存和 AMD Ryzen 9 5900X CPU 的计算机上完成。训练过程直至损失函数不再减少为止，训练期间通过验证集选取最佳权重。测试阶段不对输入作任何处理。

### 2.2　性能衡量手段

分割性能通过定性定量的手段进行评估,一方面可视化地对比分割结果，另一方面通过平均准确率(mAcc)和平均交并比(mIoU)进行数值指标分析。mAcc 衡量目标图像像素在所有语义类别上正确归类的平均概率：

$$mAcc = \frac{1}{N}\sum_{i=1}^{N}\frac{TP_i}{TP_i + FN_i} \tag{6}$$

其中，N 为类别总数，这里 N 取 9。$TP_i$ 表示正确预测为第 i 类的像素个数，即真阳性，$FN_i$ 表示被错误预测为非 i 类的像素个数，即假阴性。mIoU 衡量所有类别上的预测分割和真值标签的平均重叠率：

$$mIoU = \frac{1}{N}\sum_{i=1}^{N}\frac{TP_i}{TP_i + FP_i + FN_i} \tag{7}$$

其中，$FP_i$ 表示被错误预测为第 i 类的像素个数，即假阳性。两个指标的数值大小同分割性能正相关。

### 2.3　实验结果分析

所提网络的分割性能通过在 MFNet 测试集上进行实验验证，相关比较方法涉及当前前沿的 RGB-T 分割模型，所有数据来源于对应文章及其开源代码。表 1 和图 4 分别提供了定量比较结果和昼夜图像序列的可视化对比结果供参考。

表 1　在 MFNet 测试集上的对比结果(黑体值为最佳结果，—表示未提供项)

Tab. 1　Comparison results on the mfnet test dataset (bold values are the best results, - means no items are provided)

| Methods | Car | | Person | | Bike | | Curve | | Car Stop | | Guardrail | | Color Cone | | Bump | | mAcc | mIoU |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Acc | IoU | Acc | IoU | Acc | IoU | Acc | IoU | Acc | IoU | Acc | IoU | Acc | IoU | Acc | IoU | | |
| MFNet[7] | 77.2 | 65.9 | 67.0 | 58.9 | 53.9 | 42.9 | 36.2 | 29.9 | 12.5 | 9.9 | 0.1 | 0.0 | 30.3 | 25.2 | 30.0 | 27.7 | 45.1 | 39.7 |
| RTFNet[9] | 93.0 | 87.4 | 79.3 | 70.3 | **76.8** | 62.7 | 60.7 | 45.3 | **38.5** | 29.8 | 0.0 | 0.0 | 45.5 | 29.1 | **74.7** | 55.7 | 63.1 | 53.2 |
| AFNet[11] | 91.2 | 86.0 | 76.3 | 67.4 | 72.8 | 62.0 | 49.8 | 43.0 | 35.3 | 28.9 | 24.5 | 4.6 | 50.1 | 44.9 | 61.0 | **56.6** | 62.2 | 54.6 |
| MLFNet[12] | — | 82.3 | — | 68.1 | — | **67.3** | — | 27.3 | — | **30.4** | — | **15.7** | — | **55.6** | — | 40.1 | — | 53.8 |
| MMNet[18] | — | 83.9 | — | 69.3 | — | 59.0 | — | 43.2 | — | 24.7 | — | 4.6 | — | 42.2 | — | 50.7 | 62.7 | 52.8 |
| 本文方法 | **94.1** | **88.2** | **80.8** | **70.8** | 76.2 | 62.4 | **61.5** | **47.0** | 26.8 | 19.9 | **32.9** | 4.3 | **51.4** | 45.5 | 69.1 | 55.5 | **65.8** | **54.7** |

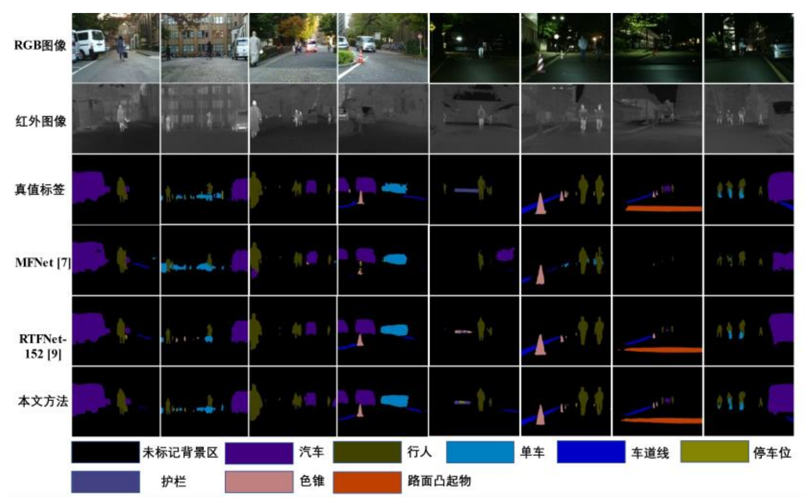



图 4　在 MFNet 测试集上的分割结果对比

Fig. 4　Comparison of segmentation results on the mfnet test dataset

据表 1 可知，所提分割网络在 mAcc 和 mIoU 两个指标上都取得了最佳值。具体而言，汽车和行人的语义归类拥有全面的最高指标，这很可能得益于嵌套形式的巢式连接和深层差异特征融合策略的共同作用，前者使得大尺度目标和易识别对象的学习能力更强，后者能够增强具有显著特征差异的目标的深层语义表达，在夜晚具有相对最大特征模态差异的汽车和行人类受此益处最多。对于车道线，由于其白色反光特性在夜晚有着稍逊于热辐信息的成像优势，一定程度上增强了自身特征优势。相对而言，单车类由于在多个场景密集扎堆，聚簇形式的单车结合体在稠密中间层可能被过拟合训练，削弱了单个单车的尺度分割优势，分割精度因而未能取得最好。而小尺度对象的色锥则很可能受此益处，这点 MFNet 和 RTFNet 可以证明，后两者网络模型未有桥接编码器和解码器的信息流通渠道，它们对于小尺度对象的特征学习能力不够。而 AFNet 和 MLFNet 分别由于共注意力融合和编码特征多级跳跃的优势在一定程度上促进了各个尺度对象的特征处理能力，各自都有着出色的分割能力。对于其他类别，护栏和停车位在测试集中的样本数过少(护栏在 393 对图像中仅有 4 对出现)，各个模型的分割情况都表现欠佳，尤其是 MFNet 和 RTFNet，这可能由于这两类本不充足的特征信息在缺乏特征复用或调节的网络训练过程中丢失过多所致。更多的细节差异可从图 4 观察比较，仅以图 4 中第 2 列和末列为例，单车类具有同真值最接近的分割情况。

为进一步探究模型的分割效能，表 2 列出了在 MFNet 测试集上单独对所有白天图像和夜间图像的实验比较结果。

表 2　昼夜图像序列对比结果

Tab. 2　Comparison results of day and night image sequences

| Methods | 白天 | | 夜晚 | |
|---|---|---|---|---|
| | mAcc | mIoU | mAcc | mIoU |
| MFNet[7] | 42.6 | 36.1 | 41.4 | 36.8 |
| RTFNet[9] | 60.0 | 45.8 | 60.7 | 54.8 |
| AFNet[11] | 54.5 | 48.1 | 60.2 | 53.8 |
| MLFNet[12] | — | 45.6 | — | 54.9 |
| 本文方法 | 58.0 | 47.0 | 64.6 | 55.7 |

由表 2 可知，所有方法均在夜晚取得了更好的分割性能，这可能是因为在光照充足的条件下，RGB 图像已包含易于分割的丰富细节信息，热辐射信息的融入会给部分优势特征造成对冲，削弱它们的语义表现。而在夜间，两种模态特征存在更大的语义鸿沟，这时候红外信息的融入更易于提高语义划分结果。

对比昼夜测试序列结果，本文方法在夜晚场景具有更好的平均准确度和平均交并比，这从侧面佐证了所提深层差异特征融合策略能够充分整合 RGB 和红外图像特征，因为红外图像天然在夜间具有成像优势，这时候两者的特征差异表现得更加突出。

### 2.3.1 编码器骨干网络变体

DenseNet 结构的不同变体作为编码器骨干网络会带来不同的分割性能。为探究 DenseNet 变体结构对分割性能的影响，在只改变骨干网络变体的条件下重新进行训练，直至损失函数不再减少为止。图 5 显示了不同变体在 MFNet 测试集上的表现情况。

图 5 中 mFPS 表示在测试集上的平均每秒分割帧数，为同分割指标值的增长方向保持一致，实际以 mFPS 的倒数绘线。由图可知，随着 DenseNet 结构变体的复杂度增加，所提网络在准确率和交并比两个分割指标上均呈递增趋势，相比之下，对应的平均分割每帧图像所消耗时间可近似视为仅同网络层数正相关。推测此种原因在于多层架构由于参数量的提升会具备更强的分割学习能力，但网络推理速度基本只受网络深度影响。

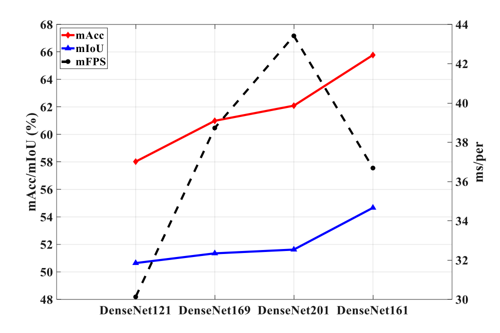



图 5　DenseNet 变体结构对分割性能影响

Fig. 5　Impact of densenet variant structure on segmentation performance

### 2.3.2 编码特征复用方式

在所提模型中，编码器和解码器间通过嵌套形式的上采样和中间层相连，这种巢式连接使得编码器的深浅特征能够以一种非常密集的形式得到复用。为了验证该做法的有效性，本部分去掉编码部分和解码部分间的所有信息复用路径，只保留编码器末尾层同解码器相连，这种结构简称为 U 型直连。同样，在 U 型结构间增加跳跃连接，用于将编码器的同层级特征传递到对应的解码器重构层，这种结构简称为同层跳跃连接。以 DenseNet-161 骨干网络为基准，保证其他条件不做变动，重训练网络直至收敛。表 3 为在 MFNet 测试集上的分割对比结果。

表 3　编码特征复用方式

Tab. 3　Multiplexing method of encoding features

| 方法 | U 型直连 | 同层跳跃连接 | 巢式连接 |
|---|---|---|---|
| mAcc | 56.1 | 67.6 | 65.8 |
| mIoU | 45.8 | 52.9 | 54.7 |

据表 3 可知，当解码器未复用编码特征时，网络分割性能急剧衰退。当通过长连接复用同尺度特征时，分割性能得到提升，尤其是准确率。而当多尺度深浅特征通过巢式连接复用时，网络的准确分割覆盖率进一步得到提升，但是单像素分割精确度略有退化。总而言之，复用编码特征会极大地影响分割性能，对深浅特征进行密集复用能最有效地提高平均交并比，但会略微削弱准确率的提高，这可能是稠密中间连接路径会对部分场景造成分割过拟合。

### 2.3.3 深层特征融合策略

为验证深层差异特征融合策略的有效性，本部分对比了两种融合策略：Transformer[19]中的自相似性融合单元和基于像素差异性的互补融合。前一策略聚焦于特征图自身各像素位置在空间位置上的相关性，是一种类似于位置注意力的融合机制，而后者关注多模态特征在像素层面上的语义相关性。相比之下，本文所提融合策略关注 RGB 和红外特征图在向量特征间的语义相关性。表 4 为这三种融合策略在 MFNet 测试集上的消融实验结果。

表 4　深层特征融合策略对比实验结果

Tab. 4　Experimental results of deep feature fusion strategies

| 方法 | 自相似融合策略 | 基于像素差异性的融合 | 基于特征差异性的融合 |
|---|---|---|---|
| mAcc | 63.7 | 65.3 | 65.8 |
| mIoU | 53.9 | 53.8 | 54.7 |

据表 4 可知，所提融合策略能提供给 RGB 和红外深层特征最佳的融合指导意义。这是因为，在多模态特征融合中，自相似融合策略忽视了相异图像特征的表达，而基于像素差异性的融合只关注局部的特征相关性，它们在整合有性质差异的多模态图像的高维特征上存在局限。总而言之，在高维抽象特征融合上，对于成像机制相异的多模态对象而言，通过挖掘它们各自不同的特征，并进行针对性的特征级上的弥

补融合能够得到具有更健壮语义表达的融合特征。

## 3　结束语

本文设计了一种融合 RGB 和红外图像深层差异特征的巢式语义分割网络，该模型考虑到来自不同编码尺度的特征具有各个层面的语义表示，通过构建嵌套形式的中间路径实现高效的深浅特征密集复用，同时为增强 RGB 和红外图像高维抽象特征的语义表达能力，通过设计深层差异特征融合策略实现特征互补增强。与前沿网络模型在公共数据集上的对比实验表明，所提模型在分割性能上具有优越性，并且消融实验证明了特征密集复用和深层差异特征融合策略的有效性。

在未来的工作中，拟聚焦于差异特征融合策略和注意力机制相结合的优化，以期提高对复杂对象的分割准确度。同时考虑将 RGB-T 分割网络泛化迁移到能够适用于其他多模态图像的语义分割领域。

## 参考文献：


[1] Yang M, Yu K, Zhang C, *et al.* Denseaspp for semantic segmentation in street scenes [C]// Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 3684-3692.

[2] 南丽丽，邓小英. 几何距优化质心结合隶属度约束 RFCM 的脑 MRI 图像分割算法 [J]. 计算机应用研究, 2019, 36 (11): 5. (Nan Lili, Deng Xiaoying. Brain MRI image segmentation algorithm based on geometric distance optimized centroid and membership constrained RFCM [J]. Application Research of Computers, 2019, 36 (11): 5.)

[3] Yu C, Liu Z, Liu X, *et al.* DS-SLAM: A Semantic Visual SLAM towards Dynamic Environments [J]. 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) , 2018.

[4] 朱文鹏，陈莉，张永新. 基于引导滤波和快速共现滤波的红外和可见光图像融合 [J]. 计算机应用研究, 2021. (Zhu Wenpeng, Chen Li, Zhang Yongxin. Infrared and visible image fusion based on guided filtering and fast co-occurrence filtering [J]. Application Research of Computers, 2021.)

[5] Wu X, Wu Z, Guo H, *et al.* DANNet: A One-Stage Domain Adaptation Network for Unsupervised Nighttime Semantic Segmentation. 2021.

[6] Jian L, Yang X, Liu Z, *et al.* SEDRFuse: A Symmetric Encoder-Decoder with Residual Block Network for Infrared and Visible Image Fusion [J]. IEEE Transactions on Instrumentation and Measurement, 2020, PP (99): 1-1.

[7] Ha Q, Watanabe K, Karasawa T, *et al.* MFNet: Towards real-time semantic segmentation for autonomous vehicles with multi-spectral scenes [C]// 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) . IEEE, 2017.

[8] Hazirbas C, Ma L, Domokos C, *et al.* Fusenet: Incorporating depth into semantic segmentation via fusion-based cnn architecture [C]// Asian conference on computer vision. Springer, Cham, 2016: 213-228.

[9] Sun Y, Zuo W, Liu M. RTFNet: RGB-Thermal Fusion Network for Semantic Segmentation of Urban Scenes [J]. IEEE Robotics and Automation Letters, 2019: 2576-2583.

[10] He K, Zhang X, Ren S, *et al.* Deep Residual Learning for Image Recognition [J]. IEEE, 2016.

[11] Xu J, Lu K, H Wang. Attention Fusion Network for Multi-spectral Semantic Segmentation [J]. Pattern Recognition Letters, 2021 (4) .

[12] Guo, Z. , Li, X. , Xu, Q. , Sun, Z. : Robust semantic segmentation based on rgb-thermal in variable lighting scenes. Measurement 186, 110176 (2021) .

[13] Zhou Z, Siddiquee M, Tajbakhsh N, *et al.* UNet+: A Nested U-Net Architecture for Medical Image Segmentation [J]. 2018.

[14] Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation [C]// International Conference on Medical image computing and computer-assisted intervention. Springer, Cham, 2015: 234-241.

[15] Huang G, Liu Z, Laurens V, *et al.* Densely Connected Convolutional Networks [J]. IEEE Computer Society, 2016.

[16] Milletari F, Navab N, Ahmadi S A. V-net: Fully convolutional neural networks for volumetric medical image segmentation [C]// 2016 fourth international conference on 3D vision (3DV) . IEEE, 2016: 565-571.

[17] Glorot X, Bengio Y. Understanding the difficulty of training deep feedforward neural networks [C]// Proceedings of the thirteenth international conference on artificial intelligence and statistics. JMLR Workshop and Conference Proceedings, 2010: 249-256.

[18] Lan, X. , Gu, X. & Gu, X. MMNet: Multi-modal multi-stage network for RGB-T image semantic segmentation. Appl Intell (2021) . https://doi.org/10. 1007/s10489-021-02687-7.

[19] Vaswani A, Shazeer N, Parmar N, *et al.* Attention Is All You Need [J]. arXiv, 2017.